In silico local structure approach: A case study on Outer Membrane Proteins.

Juliette Martin, Alexandre de Brevern, Anne-Claude Camproux

In silico local structure approach: a case study on Outer Membrane Proteins

Juliette Martin [a,b], Alexandre G. de Brevern [a], Anne-Claude Camproux [a]

Addresses: [a] INSERM UMR-S 726/Université Denis Diderot Paris 7, Equipe de Bioinformatique Génomique et Moléculaire, F-75005 Paris, [b] INRA UR1077, Unité Mathématique Informatique et Génome, F-78350 Jouy-en-Josas.

contact author: juliette.martin@jouy.inra.fr, tel: 00 33 1 44 27 99 24, fax: 00 33 1 43 26 38 30

Abstract

The detection of Outer Membrane Proteins (OMP) in whole genomes is an actual question, their sequence characteristics have thus been intensively studied. This class of protein displays a common β-barrel architecture, formed by adjacent anti-parallel strands. However, due to the lack of available structures, few structural studies have been made on this class of proteins. Here we propose a novel OMP local structure investigation, based on a structural alphabet approach, i. e., the decomposition of 3D structures using a library of four-residue protein fragments. The optimal decomposition of structures using hidden Markov models results in a specific structural alphabet of 20 fragments, six of them dedicated to the decomposition of β-strands. This optimal alphabet, called SA20-OMP, is analyzed in details, in terms of local structures and transitions between fragments. It highlights a particular and strong organization of β-strands as series of regular canonical structural fragments. The comparison with alphabets learned on globular structures indicates that the internal organization of OMP structures is more constrained than in globular structures. The analysis of OMP structures using SA20-OMP reveals some recurrent structural patterns. The preferred location of fragments in the distinct regions of the membrane is investigated. The study of pairwise specificity of fragments reveals that some contacts between structural fragments in β-sheets are clearly favored whereas others are avoided. This contact specificity is stronger in OMP than in globular structures. Moreover, SA20-

OMP also captured sequential information. This can be integrated in a scoring function for structural model ranking with very promising results.

Introduction

According to a recent study, about 25% of all proteins of a genome are related to the membrane [1]. Membrane proteins fall into two classes: most of them span the membrane with α-helices, and the others with β-strands. α-membrane proteins are mostly found in cytoplasmic membranes while β-membrane proteins are exclusively found in the outer membranes of bacteria, mitochondria and chloroplasts. The latter are thus called Outer Membrane Proteins (OMPs). They represent approximately 2-3% of the genes in Gram-negative bacterial genomes [2]. All OMPs have in common a characteristic β-barrel architecture, with 8 to 22 β-strands spanning the membrane. Some OMPs are biologically active only in trimeric form [3]. Due to experimental difficulty such as denaturation, the number of OMP structures available in the Protein Data Bank [4] (PDB) is limited. 45 different OMP structures are currently listed in the database "Membrane proteins of known 3D structures" [5].

The available structures of OMPs obey some construction principles as stated by Schulz [3, 6]: e. g., all β-strands are anti-parallel and locally connected to their next neighbors; both the N- and C-termini are at the periplasmic end; the external connections are long loops, whereas periplasmic connections are shorts. Despite these common structural features, the biological functions of OMPs are diverse, e. g., passive transport through the membrane or enzymatic activity [3]. OMPs have biomedical interest because they contribute to the pathogenicity of bacteria [6] and are involved in antibiotic resistance [7].

Since they are dedicated to the particular environment of a lipidic bilayer, OMPs have specific amino-acid composition [8] and original properties. For example, the surrounding hydrophobicity of aromatic and non-polar amino acid residues is lower in OMPs than in globular proteins [9]. The identification of β-barrel is an important problem. The specific composition of OMPs has been used to discriminate β-barrels from other proteins [10, 11, 12]. Several groups use learning methods for the identification of β-barrel in entire proteomes and the prediction of transmembrane segments. The methodologies includes hidden Markov models [13, 14, 15, 16], neural networks [17, 18, 19, 20], support vector machine [21, 22] and nearest neighbors algorithms [23]. An evaluation of different methods for predicting the topology

of β-barrel can be found in [24] and a comparison of learning methods for discrimination is presented in [25]. Most recent developments include the use of grammars, i. e., a generalization of hidden Markov models [26] to predict interstrand residue interactions.

The structural study of OMPs, however, has been less addressed. In their study, Seshadri and coworkers [27] showed that the amino-acid conservation is greater at the trimeric interface. Wimley [2] computed hydrophobicity profiles along the barrel axis and analyzed the relative amino-acid abundance on the internal and external sides of the membrane. Gromiha and Suwa [9] analyzed various amino acid features in globular and OMP structures. Jackups and Liang investigated in a very elegant manner the amino-acid propensities in various regions of OMP structures [28] and the specific motifs and antimotifs of OMP strands [29]. However, the question of an eventual intrinsic difference in terms of local structure between OMPs and globular β-proteins remains an open question.

We propose an investigation of OMPs in terms of local structures. This investigation is made through the learning of a structural alphabet. A structural alphabet is a collection of local protein 3D fragments that allows the decomposition of protein structures into series of short structural fragments [30, 31, 32, 33]. In particular, this provides a more precise description than the classical secondary structures. Structural alphabets constitute a powerful approach for structure approximation [33, 34], structure mining [35, 36, 37] and a promising tool for structure prediction [38, 39]. In a previous work, a structural alphabet was identified from a set of globular proteins using Hidden Markov Model (HMM) [31, 34]. An advantage of using HMM is that learning takes into account both the structure of the fragments and the connections between them in the protein structures. This allows the identification of structural letters with relatively similar structures but concatenated in different ways to form different longer fragments. This alphabet, called SA27, is made of 27 structural letters of four residues, with specific transition rules between the letters. It has been shown on globular proteins, that SA27 offers an excellent local approximation of 3D structures. In this framework, we propose to apply a similar approach to analyze OMP structures.

In this paper, we describe the novel structural alphabet learned on OMP structures and highlight the interest of such an approach. This alphabet allows the description of OMP structures as series of four-residue overlapping fragments. The transitions between fragments, learned thanks to a HMM, are governed by highly specific transition rules. We address the critical question of the number of clusters, i. e., structural letters of the new alphabet, and show the differences observed when using a similar

approach on globular proteins. In the same way, we analyze the topology of β-strands in OMPs, both locally and pairwise. Finally, we present a novel approach to select structural model of OMPs within a set of different potential structures submitted at CASP3 meeting, using the sequence-structure relationship of our specific alphabet.

Material and Methods

Data

Two different data sets are used in this study. The first one is composed of 17 outer membrane proteins and is named *OMPset*. These X-ray structures have a resolution better than 2.6 Å and a sequence identity less than 26%. It is the same data bank used by Jackups and Liang [28]. Nonetheless, we have excluded two proteins (PDB code [4]: 1ek9 and 7ahl) for which several distinct chains contribute to the barrel. *OMPset* corresponds to 6143 overlapping four-residue fragments and 6239 residues (32 protein fragments). Secondary structures are assigned by a classical method, STRIDE [40], and a recent method, KAKSI [41]. The 8 states assigned by STRIDE are reduced into three states as follows: E=β, (H, G, I)=α, others=coil. The secondary structure content of this set is 5% α-helix, 62% β-strand and 33% coil using the STRIDE definition [40]. The β-sheet assignment in KAKSI is done by checking distances and angles in two sliding windows along the protein backbone. A contact matrix is generated, containing the positions of the sliding windows where the criteria assignment were fulfilled. We have used this matrix to analyze the neighboring residues in β-sheets.

The second data bank called *GBset* is composed of globular protein structures with a high level of β-sheet, a resolution better than 2.5 Å and less than 25% of sequence identity. A pre-compiled list of protein chains was retrieved from the PISCES website [42]. It corresponds to 3925 structures (cullpdb_pc25_res2.5_R1.0_d060903_chains3925). A minimal threshold of 56% of β-sheet content per protein has been selected, leading finally to a data set of 89 proteins (β-sheet assignment by KAKSI [41] software). Structures were filtered on the RPBS web-site [43], to remove chains with missing coordinates or poor backbone geometry. In the same way, membrane proteins were excluded. 38 proteins were finally selected. *GBset* is composed of 4788 overlapping four-residue fragments and 4902 residues, with a secondary structure content consisting in 5% α-helix, 58% β-strand and 37% coil.

In order to analyze the relevance of the sequence-structure relationship of our alphabet, we develop a novel approach to select pertinent structural models of OMPs structures. For this purpose, we use structural models of Omp32, the anion-selective porin from *Comamonas acidovorans*, that was submitted at Critical Assessment of Protein Structure Prediction 3 [44]. The structural models, obtained by different scientific teams, using various strategies, were downloaded from the Protein Model Database [45]. The models were compared with the crystallographic structure of Omp32, deposited since in the PDB with id 1e54 [46].

Hidden Markov model-based structural alphabet

A structural alphabet is a collection of short protein fragments, used to approximate protein 3D structures [38, 31, 34]. The use of hidden Markov models (HMM) [47] to determine structural alphabet and the learning procedure are detailed in [31, 34]. The structural alphabet specific of OMP structures is obtained as done previously with globular proteins [31, 34] using *OMPset*. The underlying hypothesis is that all structures share the same canonical shapes of four-residue and the same logic of assembly [34].

Protein structures are cut into overlapping four-residue fragments. Each four-residue fragment is described by a vector of four distances between its Cα:

$$
\begin{aligned}
d_1 &= \|\overrightarrow{C_{\alpha 1}C_{\alpha 3}}\| \\
d_2 &= \|\overrightarrow{C_{\alpha 1}C\alpha_4}\| \\
d_3 &= \|\overrightarrow{C_{\alpha 2}C_{\alpha 4}}\| \\
d_4 &= \frac{\overrightarrow{C_{\alpha 1}C_{\alpha 2}} \wedge \overrightarrow{C_{\alpha 2}C_{\alpha 3}}}{\|\overrightarrow{C_{\alpha 1}C_{\alpha 2}} \wedge \overrightarrow{C_{\alpha 2}C_{\alpha 3}}\|}^{\top} \overrightarrow{C_{\alpha 3}C_{\alpha 4}}
\end{aligned}
$$

where d_1, d_2, d_3 represent euclidean distances between non successive Cα, and d_4, the signed descriptor between $C_{\alpha 4}$ and the plane P formed by $C_{\alpha 1}$, $C_{\alpha 2}$ and $C_{\alpha 3}$. d_1 and d_3 describe respectively the extension of the N-terminal and C-terminal part of the fragment. d_2 describes its total extension. The absolute value of d_4 is a measure of the fragment volume, i.e., a flat fragment having a d_4 close to zero. The sign of d_4 allows the distinction between the fragment and its mirror image: a positive (respectively negative) value indicates that $C_{\alpha 4}$ is located above (respectively below) the plane P, in the trigonometric sense. These four descriptors are modeled as a mixture of four-dimensional multi-normal distribution.

Then, a hidden Markov model is trained on these data. The parameters of such a model are:

- the number of fragment clusters n, i.e., the number of hidden states,
- for each hidden state $1 \leq i \leq n$, a four dimension multi-normal density of parameters θ_i describing the means and variance-covariance matrix of the four descriptors d_1, d_2, d_3, d_4,
- the transition probabilities between hidden states corresponding to the Markov process, forming a matrix transition with element $P[i, j]$ being the probability p_{ij} to transit from letter $1 \leq i \leq n$ to letter $1 \leq j \leq n$.

Parameters are estimated from the data set using the expectation-maximization (EM) algorithm [47], as explained in [34]. The resulting model is a structural alphabet made of n fragment clusters, named structural letters, and the transition rules between them.

The optimal size of the structural alphabet, i.e., the number n of structural letters, is chosen according to the Bayesian Information Criterion (BIC) [48]. The BIC is a penalized log-likelihood criterion used to select the model that ensures the best compromise between the fit to the data and a correct parameter estimation.

Structural alphabet analysis

To measure the structural variability within a cluster (i. e., a structural letter), 50 pairs of fragments belonging to that cluster, randomly chosen, are superimposed to compute the Cα root mean square deviation, named rmsd_w (w for *within*). The structural variability between two different structural letters, rmsd_b, is computed in the same way.

Representative fragments of structural letters are extracted from the data set using the distance descriptors. All fragments encoded by a given structural letter are considered. The fragment whose distance vector is the closest to the mean descriptors (assessed by the sum of the absolute deviation) is chosen as the representative fragment of this letter.

The number of equivalent output of a structural letter, N_{eq}^o, is given by:

$$N_{eq}^o(i) = e^{H(i)},$$

with $H(i) = -\sum_{1 \leq j \leq n} p_{ij} ln(p_{ij})$ where p_{ij} is the probability, for state i, to transit to state j. It is derived from the Shannon entropy. For a structural alphabet of size n, the N_{eq}^o varies between 1, for a

state leading to only one other, to n for a letter leading to all the others with equal probabilities. The number of input equivalent, N_{eq}^{i}, is computed similarly.

The average number of repetition of a structural letter, ANR, is given by:

$$ANR(i) = \frac{1}{1 - p_{ii}},$$

where p_{ii} the probability of self transition for state i. For $p_{ii} = 0$, the ANR is 1.

OMP structure encoding

3D structures are encoded in terms of structural letters using the Viterbi algorithm [47] that computes the most probable sequence of structural letters given the fragment descriptors. A 3D structure of a protein of N residues is then encoded as a sequence of $N - 3$ structural letters. The Viterbi algorithm also provides the associated log-likelihood [49].

OMP structure analysis

The correspondence between Cα and structural letter encoding is made between one structural letter and its third Cα, in particular for secondary structure assignation.

The specific localization of structural letters in various parts of OMP structures is explored using the database of Lomize et al [50]: Orientation of Proteins in Membranes database. This database contains the coordinates of membrane protein structures positioned in artificial membranes. The positions are obtained by minimizing the transfer energy from water to the membrane core [51]. Membranes are materialized by a layer of oxygen atoms (extra-cellular side) and one of nitrogen atoms (periplasmic side). Each Cα of a structure is then annotated as extracellular, membrane-spanning or periplasmic. Using a definition similar to Jackups and Liang [28], we further detail the annotation into five regions. The core (C) region encompasses all residues within 6.5 Å of the barrel center. The periplasmic and extracellular head-group regions, (P) and (E), encompass the space remaining between the core and the membrane limit. The periplasmic extracellular cap regions, (p) and (e), are 7 Å thick regions outside the membrane. Frequency of structural letters are computed separately in the five regions, with the correspondence between one structural letter and its third Cα. A χ^2-test is used to assess the influence of

the region on the structural letter composition. The over and under-representations of structural letters in each region are measured by Z-scores:

$$Z_{r,i} = \frac{N_{r,i}^{obs} - N_{r,i}^{exp}}{\sqrt{N_{r,i}^{exp}}},$$

where $N_{r,i}^{obs}$ is the observed frequency of letter i in region r and $N_{r,i}^{exp}$ is its expected frequency if the letter distribution is similar in every region: $N_{r,i}^{exp} = N_i \times F_r$ with N_i the frequency of structural letter i and F_r, the proportion of structural letters involved in region r. The threshold for Z-score significance is corrected using the Bonferroni method to take into account multiple tests. The non-parametric approach described by Jackups and Liang is also used [28]. Using a null model of exhaustive permutations, it allows the explicit computation of p-values.

Neighboring Cα in β-strands are analyzed using KAKSI. Statistics are then derived from the pairs of neighboring structural letters. The Z-score for a neighboring ij pair is given by:

$$Z_{ij} = \frac{N_{ij}^{obs} - N_{ij}^{exp}}{\sqrt{N_{ij}^{exp}}}$$

where N_{ij}^{obs} is the observed frequency of the ij pair and N_{ij}^{exp} is its expected frequency if pairs are random, given by $N_{ij}^{exp} = \frac{N_i \times N_j}{N_{tot}} \times F$ where N_i denotes the frequency of structural letter i in the neighbors, N_{tot} denotes the number of all structural letters in the neighbors, $F = 2$ if $i \neq j$ and 1 if $i = j$. F factor allows to make no distinction between ij and ji pairs. Note that in anti-parallel β-sheets, two structural letters are neighbors if the second Cα is paired with the third Cα of the other.

Sequence specificity

Once structures are encoded into structural letters, frequencies of amino-acids at each position of the structural letters are computed. The Z-score, for amino-acid a, at position $1 \leq p \leq 4$ of the structural letter $1 \leq i \leq n$ is given by:

$$Z_{a,p,i} = \frac{N_{a,p,i}^{obs} - N_{a,p,i}^{exp}}{\sqrt{N_{a,p,i}^{exp}}}$$

with $N_{a,p,i}^{obs}$, the observed frequency of amino-acid a at position p of the structural letter i, and $N_{a,p,i}^{exp}$, the expected frequency if the amino-acid repartition in structural letters is random. $N_{a,p,i}^{exp} = \frac{N_{p,a} \times N_i}{N_{tot}}$,

where $N_{p,a}$ denotes the frequency of amino-acid a at position p of every structural letters, N_i denotes the frequency of structural letter i, and N_{tot}, the total number of structural letters. A positive (respectively negative) Z-score indicates that amino-acid a is over-represented (respectively under-represented) at position p of the letter i.

Sequence-to-structure adequacy

The new structural alphabet is used to assess the sequence-to-structure adequacy of the quality of structural models. Hence, 47 structural models of Omp32 submitted to CASP3 meeting by different predictor groups, and the true structure of Omp32 (PDBcode 1e54), are encoded into structural letters and receive a score defined by:

$$S = \frac{1}{N_F} \sum_{i=1}^{F} log_2 \frac{P(a_1 a_2 a_3 a_4 \mid s_i)}{P_{rand}(a_1 a_2 a_3 a_4)}$$

where N_F denotes the number of overlapping four residue fragments in the model. $P(a_1 a_2 a_3 a_4 \mid s_i)$ is the probability of observing the sequence $a_1 a_2 a_3 a_4$ in the fragment i encoded by the structural letter s_i. It is given, under independence assumption, by: $\Pi_{p=1}^{4} P(a_p \mid s_i)$ where $P(a_p \mid s_i)$ is the probability of finding residue a_p in position p of structural letter s_i. These probabilities are estimated by the observed frequencies, using a pseudo-count to avoid the problem of zero probability. $P_{rand}(a_1 a_2 a_3 a_4)$ represents the probability of the amino-acid sequence $a_1 a_2 a_3 a_4$ in OMP sequences under a random, memoryless model where the sequence is formed by picking amino-acids according to their frequency in OMP sequences. It is estimated in the same manner as $P(a_1 a_2 a_3 a_4 \mid s_i)$, using the frequencies of amino-acids in the *OMPset*. This score allows to measure the agreement between the sequence and the local structure of the model,defined by structural letters. A positive (resp. negative) score indicates that the sequence is in good (resp. poor) agreement with the local structure.

A similar score can be defined, using the 3 classes of secondary structure (helix/strand/coil), instead of the n structural letters. In that case, the local structure of a four residue fragment is defined by the secondary structure state of its third residue, defined by STRIDE.

To evaluate the relevance of this approach, the 47 structural models are compared to the X-ray structure using Cα rmsd.

Results

Data description

A non-redundant data set of 17 OMP structures, called *OMPset*, is used in this study (see Table I). Some example structures are shown on Figure 1a. The secondary structure of the *OMPset* is analyzed with two software: the classical method STRIDE [40] and a recent method, KAKSI [41]. The later method is used to study the pairing between β-strands. These 17 protein chains cover the classical size range of the OMP, with 8 to 22 strands forming the barrel. The mean length of β-strands assigned by STRIDE greatly varies from 13 to 26 residues. This value refers to the whole length of β-strands; the length of β-strands actually spanning the membrane is shorter. Following Lomize et al approach [50], only nine residues are sufficient for the transmembrane region. Few β-strand residues are found in the periplasmic regions, but long β-strands expanding from the membrane in the extracellular regions are common. Most of the residues in transmembrane region (about 80% using STRIDE assignment) are associated to β-strands. As expected, no correlation can be found between the mean length of β-strands and the barrel size. In the same way, no correlation can be found between strand length and the biological unit, i. e., monomeric or tetrameric.

Analysis of the OMP-specific alphabet

Optimality of the alphabet

A critical question when generating structural alphabets is the choice of the number of clusters, or structural letters, forming the alphabet. Following the previous work [34], the Bayesian Information Criterion (BIC) [48] is used to determine the optimal size of the structural alphabet. The BIC criterion is the data likelihood penalized by a term related to the number of parameters and to the amount of data; it ensures the best compromise between the fit to the data and the number of parameters. As expected, its optimum depends on the size of the learning data set, i. e., the number of available structures. Structural alphabets learned on respectively 40 and 60% of the *OMPset* have their respective optima at 13 and 17 structural letters. Using the complete *OMPset*, the maximum BIC value is obtained for a size of 20 structural letters (see supplementary data). This alphabet, designated as SA20-OMP, corresponds to an

optimal description of available structures and has many interesting features, has shown further.

To check the consistency of SA20-OMP, models with 20 structural letters are estimated on the two independent subsets previously mentioned, containing respectively 40 and 60% of the data. These models, noted SA20-OMP$_{40\%}$ and SA20-OMP$_{60\%}$ are compared to SA20-OMP. The similarity of these alphabets is assessed by the similarity of the descriptors of structural letters. 16 structural letters of SA20-OMP$_{40\%}$ are identical to structural letters of SA20-OMP. These 16 structural letters encompass 87% of the fragments. In the same way, 17 structural letters of SA20-OMP$_{60\%}$, accounting for 91% of the fragments, are identical to structural letters of SA20-OMP. The stability of SA20-OMP is further analyzed using a jackknife method: each protein is removed from the *OMPset* and a new alphabet is learned. The examination of the mean descriptors of the 17 resulting alphabets indicates that they are 90 to 100% similar to SA20-OMP.

Description of the optimal alphabet

The 20 structural letters of SA20-OMP are described in Table II, by their mean descriptors d_1, d_2, d_3 and d_4. Structural letters are symbolized by capitalized letters ranging from A to T (in italic in the text to avoid the confusion with amino-acids). Their relative frequency varies from 2.1% for letter T to 10.3% for letter J. The total extension of the structural letters, measured by d_2, ranges from 5.35 Å for letter A, to 10.57 Å for letter Q. For 13 letters out of 20, d_4, measuring the fragment volume, is the most variable descriptor and for the remaining 7, it is d_2. The geometric variability of the letters, assessed by the rmsd$_w$ ranging from 0.21 to 1.03 Å is low, with only four letters having rmsd$_w$ greater than 0.5 Å, and a mean rmsd$_w$ of 0.35 Å. The rmsd across distinct structural letters, rmsd$_b$ are all greater than the rmsd$_w$, except for the fuzziest letter G. The clusters are thus well defined and well separate except G that gathers variable fragments (2.3% only of the data). The overall examination of Table II indicates that three main groups emerge from the matching with β-strands: [$ANTHD$], [$ESORGFPKL$] and [$JMBICQ$].

Letters [$ANTHD$], accounting for 16.4% of the fragments, are never or rarely seen in β-strands (less than 9%). They are said *incompatible* with the β-strand conformation. These letters are characterized by short d_2, from 5.35 Å to 8.25 Å, and high absolute d_4, around 3 Å for all letters but N. Such values are characteristic of rather helical conformations. Indeed, comparison with STRIDE assignment indicates that respectively [48.4, 10.0, 40.0, 3.9, 16.2]% of letters [$ANTHD$] correspond to α-helices (data not

shown). Their rmsd_w vary between 0.25 Å for letter *A* and 0.65 Å for letter *H*.

Letters [*ESORGFPKL*], accounting for 35.2% of the fragments, are sometimes associated with β-strands and are thus said *β-compatible*. The proportion corresponding to a β assignment by KAKSI in this group varies from 18% for letter *O* to 61% for letter *L*. These letters have intermediate conformations: d_2 between 7.76 and 9.78 Å and large range of volume, with d_4 between -3.24 Å and 2.75 Å. This is the most variable group, but the rmsd_w are lower than 0.6 Å except for letters *E* and *G*. The ratio involved in β-strands clearly depends on the assignment software for two letters: 47% of fragments in letter *E* are in β-strand according to STRIDE but 21% according to KAKSI; 53% of fragments in letter *S* are in β-strand according to STRIDE but 33% according to KAKSI. The discrepancies between STRIDE and KAKSI assignments have already been observed and analyzed elsewhere [41], they correspond mainly to strand ends.

The six remaining structural letters [*JMBICQ*], accounting for 48.4% of the data, are clearly associated to β-strands (β ratio more than 89%) and are thus said *β-specific*. All these letters are very extended with d_2 ranging from 10.05 to 10.57 Å. The descriptor d_4 is always the most variable. Letters *J* and *Q* are symmetric with flat volumes: d_1 equal to d_3 and d_4 is close to zero. They are respectively the shortest and longest β-specific letter. Letters *M* and *C* are more extended at N-terminal than C-terminal end (d_2 greater than d_3); *M* is globally shorter and has flat volume whereas *C* has a mean d_4 equal to -2 Å. Letters *B* and *I* are more extended at C-terminal end than N-terminal ends, specially *I*. Letters [*JMBQ*] represent each about 10% of the data while letters [*IC*] are less frequent (about 5% of the data). The rmsd_w in this group is very low, ranging from 0.2 Å to 0.3 Å. These letters are thus very well defined.

SA20-OMP is obtained using an iterative procedure: starting from a two-state alphabet, a new state is added at each step, until 20 states. States are labeled from *A* to *T* according to their order of creation. When a new state is created, most of the time, it results from the splitting of a state from the previous step. Some state creations are more complex: a new state can result from the grouping of fragments from several parent states. An interesting feature of this procedure is the possibility to follow the genesis of SA20-OMP. Figure 2 presents a simplified view of the genesis with the main splitting events and the evolution of the corresponding mean rmsd_w. Helical letters (*ANTH*) have their common ancestor at step 5, and β-specific letters (*JMBICQ*), at step 2. The last differentiation occurs between letters *A* and *T*, with a rmsd_b equal to 0.46 Å. At the very early stage of two letters, the HMM classification results in one

extended letter, the common ancestor of the β-specific letters [*JMBICQ*], and one letter that encompass all other fragments. As expected since OMP structures are all-β proteins, the predominant feature of the classification is the distinction between strands and non-strands. At three structural letter decomposition, the β-fragments are split into two distinct clusters: extended letters [*BMJ*] and very extended letters [*CQI*]. Then, most of the divisions occur in the non β-fragments. The differentiations of new β-specific letters later occur at step 8 (letter *I* with positive volume *vs* [*CQ*] with negative volume), step 9 (*B*, almost flat, *vs* [*JM*] with negative volume), step 12 (*J*, with d_1 equal to d_3, *vs M*, more extended at N-ter) and step 16 (*C*, with big volume and more extended at N-ter *vs Q*, with small volume and symmetrical).

The usage of hidden Markov models to identify our alphabet provides a set of geometrical fragments, but also the transition probabilities between these fragments. These transition probabilities can be seen as local building rules to create OMP structures with the 20 structural letters.

Figure 3 is a graphical representation of the transition matrix of SA20-OMP, with element $P[i, j]$ being the probability p_{ij} to transit from the structural letter i to the structural letter j. For clarity, probabilities lower than 0.01 are not indicated. The transition matrix is very sparse: only 211 transition probabilities, out of 400, are greater than 0.01, and 61 are greater than 0.1. Twelve probabilities only are greater than 0.3. The number of output equivalents, N^o_{eq}, ranges from 3.95 for letter *B* to 13.65 for letter *0*, with a mean at 7.05. This is significantly less than 20, that would be obtained with all equal transitions. The number of input equivalents, N^i_{eq}, varies between 3.11 for letter *C* and 11.56 for letter *I* with a mean at 6.89. The average number of repeats, ANR, is between 1 and 1.75, with a mean at 1.19.

The examination of transition probabilities greater than 0.1 (Figure 3a) reveals a structure with roughly four groups: [*ANTH*], [*ESO*], [*DRGFPKL*] and [*JMBICQ*]. Group [*ANTH*] leads to itself and to [*ESO*]. Its mean N^o_{eq} is 6.18 and mean ANR is 1.27. Letter *N* has the highest N^i_{eq}, 11.2, indicating that it it can be reached by many structural letters. Group [*ESO*] (differentiated at step 5, see Figure 2) transits to groups [*DRGFPKL*] and [*JMBICQ*]. Its mean N^o_{eq} is 11.29, and ANR are close to 1, meaning no repetitions. Group [*DRGFPKL*] leads to [*ANTH*], [*ESO*] and [*DRGFPKL*]. N^o_{eq} values in this group are between 5.07 and 12.97, with a mean at 9.92 and ANR are between 1 and 1.17 with a mean at 1.09. The mean N^o_{eq} 9.92 and mean ANR is 1.09. Group [*JMBICQ*] leads to itself and to [*DRGFPKL*]. The mean N^o_{eq} of these β-specific letters is 4.73 and mean ANR is 1.27. To summarize, the β-incompatible group [*ANTH*] and the β-specific group [*JMBICQ*] both have low N^o_{eq} and high ANR values. They do

not communicate directly, but *via* clusters [*ESO*] and [*DRGPFKL*], that are characterized by high N^o_{eq} and low ANR and communicate together.

The information about transition rules supplements the structural description. Taking into account both information during the learning allows to separate some letters that are relatively close in geometry but concatenated in different ways to form different longer fragments. For instance in the β-specific group, letters M and B are structurally close but different transition profiles. Even the fuzziest letter G has N_{eq} and N^i_{eq} values far from 20 (respectively 10.10 and 7.76), indicating strong transition constraints. The output profile of G is different from the one of letter R which is structurally close.

Interestingly, β-specific group [*JMBICQ*] presents many very low and very high transition probabilities: height of the transitions higher than 0.3 occur within this group, indicating very constrained transitions between letters. The values of N^o_{eq} and N^i_{eq} of these letters are around 4, except C that has a N^o_{eq} equal to 8.32 and I that has a N^i_{eq} equal to 11.56. These values indicate that C and I are more connected to other states than other letters in the β-specific group with C having more output and I more input states.

In the same way, we can note that C and Q, that were differentiated at step 16 (see Figure 2), have distinct transition preference within β-specific cluster: C has high transition probability to J and B, while Q has high transition probability to M and Q. Thus the β-strands of OMPs can be described by six distinct structural letters connected with preferentially transitions.

Comparison with a globular-barrel specific alphabet

It is known that membrane proteins, due to their particular localization, have sequence composition distinct from globular proteins. However, concerning the local structure, few studies have been published [27, 2, 9, 28, 29]. To determine if OMP structures have particular local structure decomposition, a structural alphabet is learned on a set of 38 globular structures with the same content in β-sheets than *OMPset* called *GBset*. This structural alphabet, called SA20-GB, results in six structural letters specifically describing the β-strands. The examination of SA20-GB (supplementary data) indicates that globally, structural letters of SA20-OMP are more extended than those of SA20-GB: mean d_2 are respectively equal to 9.10 Å and 8.91 Å. This is more patent on the β-specific letter descriptors: their mean d_2 are 10.28 Å in SA20-OMP *versus* 10.05 Å in SA20-GB.

We use an approach of HMM comparison [47] to further analyze the differences between SA20-OMP and SA20-GB. This analysis is based on the likelihood of the structures under three different alphabets: SA20-OMP, SA20-GB and the alphabet previously introduced by Camproux et al [34]. This latter alphabet, called SA27, is composed of 27 structural letters, four of them describing α-helices and five of them describing β-strands. It was learned on a large number of globular structures and provides a satisfying local approximation. The structures of the *OMPset* and the *GBset* are encoded under the three alphabets using the Viterbi algorithm and the repartition of the resulting log-likelihoods are analyzed. The log-likelihood can be seen as the compatibility between a structure and a structural alphabet. The results are shown on Figure 4. When comparing SA20-OMP and SA20-GB, as expected, each alphabet gives higher likelihood to the structures used for learning (Figure 4a). When log-likelihoods obtained under SA27 are compared with those obtained under SA20-OMP (Figure 4b), the discrimination is clear: OMP structures have higher likelihoods under SA20-OMP and almost all GB structures have higher likelihoods under SA27. It means that *GBset* structures, which have the same secondary structure content than *OMPset* structures, are better represented by SA27 than SA20-OMP. On the contrary, no clear distinction can be seen when SA27 is compared with SA20-GB (Figure 4c): some *GBset* structures have higher log-likelihood under SA27, and some *OMPset* structures have higher likelihoods under SA20-GB. These results indicate that OMP structures have characteristic features in terms of local structures, and are better represented by a specific alphabet. This analysis shows the interest and necessity to learn a specific alphabet of OMP structures.

Analysis of OMP structures with SA20-OMP

The 17 structures used to learn SA20-OMP alphabet are analyzed with this structural alphabet. 3D structures are encoded in SA20-OMP alphabet using the Viterbi algorithm. Four structures colored according to β-specific letters of SA20-OMP are shown on Figure 1b.

First, we analyze the recurrence of patterns of structural letters in the structures. We then explore the pairwise specificity of structural letters in neighboring β-strands. Finally, we analyze the propensities of structural letters in different localizations of OMP structures.

Recurrent structural patterns in OMP structures are investigated by counting the frequency of short series of structural letters in the OMP structures encoded in SA20-OMP. Following other studies, we

consider patterns of 4 structural letters [52]. Figure 5a illustrates the 5 most frequent patterns of 4 structural letters (i. e. 7 residues). *AAAA* is a pattern seen in helices whereas the others are found in β-strands. All these patterns have rmsd lower than 0.5 Å, ranging from 0.34 Å for *AAAA* to 0.46 Å for *BMBM*. It is interesting to note that patterns *JJBM* are paired on neighboring β-strands, as shown on Figure 5b. We compare the frequencies of all 4 structural letter patterns in *OMPset* and in *GBset*. Results are presented in Figure 5c. It is clear from Figure 5c that *OMPset* contains more recurrent patterns than *GBset*. Moreover, unlike the *GBset*, recurrent patterns in OMP structures are series of β-specific letters (see Figure 5c). These preliminary results show, on few available structures, that SA20-OMP, combining structural letters and their transitions, allows to capture some precise recurrent structural fragments of 7 residues within β-strands that are specific of OMP structures. This confirms the finding presented in Figure 4: although they have similar global architecture, OMP and globular β-barrel have distinct local structural features.

3D contacts between structural letters forming the β-barrel are studied thanks to the KAKSI output that indicates the neighboring residues in β-sheets. 140 different pairs are observed in neighboring β-strands, i. e., 67% of the 210 potential pairs. The over and under-representation of these pairs is assessed by Z-score computation. 11 pairs are over-represented (*IR, LF, PI, LL, LC, JJ, MB, MI, QQ, BC* and *IC*) and 4 pairs are avoided (*JI, JQ, MM, BB, BI*, and *MC*), using a Bonferroni-corrected threshold of 3.7. Preferred and avoided pairs of β-specific letters are illustrated on Figure 6. We can note that *MM* and *BB* pairs are avoided, while *MB* pair is preferred. It can be explained by the geometric characteristics of the letters: *M* has mean d_1 equal to 7 Å and d_2 equal to 6.6 Å, *versus* respectively 6.6 and 7 Å for letter *B*. Since the pairing is anti-parallel, the configuration of *MB* is more favorable to hydrogen-bound formation than *MM* and *BB* pairs. Letters *J* and *Q* being symmetrical (i. e., d_1 equal to d_3), they form preferential pairs for same geometrical reasons. The analysis of neighboring structural letters in the *GBset* structures encoded in SA20-OMP, revealed less specificity. Five pairs only are significantly over-represented: *GC* (Z-score=4.4), *LL* (Z-score=3.7), *JJ* (Z-score=4.5), *MB* (Z-score=4.3) and *IC* (Z-score=4.7). Then, OMP and GB structures have similar pairwise propensities, but OMP structures present more preferred contacts than GB structures. The same analysis carried out on amino-acid revealed only two significantly over-represented amino-acid pairs: leucine with tyrosine (Z-score=4.7) and phenylalanine with valine (Z-score=3.7).

The propensity of structural letters is analyzed in specific regions of OMP structures. Thanks to the information provided by the database of Lomize et al [50], each Cα is assigned as included in the membrane, extra-cellular or periplasmic region. This allows the distinction between strands that span the membrane from the exterior to the interior (down) or from the interior to the exterior (up). However, no significant features were found concerning the propensities of structural letters [*JMBICQ*] in the different regions, or the particular composition of up- and down-strands. We then refine the definition of 5 different regions using similar definitions as Jackups and Liang [28], as extracellular cap (e), extracellular head group (E), core (C), periplasmic head group (P) and periplasmic cap (p). With this definition, the membrane region is now divided into 3 regions (E, C, and P). When the analysis is restricted to structural letters [*JMBICQ*], the usage of structural letters is distinct in the different regions, as assessed by a χ^2-test (p-value less than 10^{-6}). We then refine the analysis by computing the Z-scores for each structural letter in each region, and their p-values using the non-parametric approach of Jackups and Liang [28]. One Z-score is significant, but the explicit computation of p-values allows to capture more information, with five structural letters having p-values lower than 0.05. The results of this analysis are summarized in Figure 7. General tendencies emerge: structural letters I and J are predominant near the periplasmic, whereas structural letters M, B and C are predominant near the extracellular side of the membrane. The most extended letter, Q, is preferred in the core region of the membrane.

Adding sequence information to the structural alphabet

The classification of fragments to generate the OMP-specific structural alphabet is based exclusively on structural description of overlapping fragments. Nevertheless, sequential information can be extracted afterward. For all the 17 structures of the *OMPset*, the amino-acid sequence corresponding to each structural letter was collected to analyze the amino-acid propensities. Z-scores are computed as described in Material and Method section, for each position in each structural letter. All letters display significant Z-scores, which means that the structural alphabet also captured sequential information. Z-scores of β-specific letters are illustrated on Figure 8. Although structurally close, the β-specific letters have distinct similar amino-acid propensities. For example, structural letter B is characterized by a strong under-representation of glycine residue at its second position, whereas it is strongly over-represented at the second position of letter C. The sequential specificity reflects the strong constraints on the transitions

between structural letters. For instance, letter I transits to C with high probability. The third position of I display an under-representation of leucine, valine, proline, aspartic acid and lysine residues, and a high under-representation for glycine. The same propensities are found at the second position of letter C. Similarly, propensities at the third position of letter M, particularly high over-representation of hydrophobic residues isoleucine, valine and leucine, and a high under-representation of glycine, are similar to the propensities at the second position of letter B. This is related to the high transition probabilities from M to B. For the same reasons, letters J and Q, having high self-transition probability, display similar amino-acid propensities among their positions.

Amino-acid Z-scores can be used to perform a hierarchical clustering of the 20 structural letters. To compare the sequential and structural similarities of structural letters, the same clustering is made using the rmsd_b. For comparison with classical secondary structure, a representative fragment of β-strands assigned by STRIDE is included in the clustering. Both resulting classifications are shown on Figure 9. If we first consider the structural classification (Figure 9a), four main groups appear: $[ANT]$, $[OSEG]$, $[HRDF]$ and $[KCLMJPIBQ]$. The cluster formed by $[ANT]$ only contains helical letters, while the cluster $[OSEG]$ contains only β-compatible, including the highly variable letter G. The cluster $[HRDF]$ is composed of both β-incompatible (H and D) and β-compatible letters (R and F). The last and biggest cluster contains three β-compatible letters, $[KLP]$, together with the six β-specific letters $[JMBICQ]$. As expected, the representative fragment of STRIDE β-strands is found in this group; it is closer to J. The five main clusters obtained with amino-acid Z-scores (Figure 9b) are: $[QBI]$, $[JM]$, $[CR]$, $[GHNTEFKL]$, $[AOSDP]$. The representative fragment of STRIDE β-strands appears clustered with $[QBI]$. β-specific letter are grouped together, but split into two sub-clusters $[QBI]$ and $[JM]$. The latest sub-cluster $[JM]$ is closer to other letters than $[QBI]$. The β-specific letter C is close to the β-compatible letter R. Some structural letters are close both in structure and sequence: O and S, N and T, J and M, K and L, B and I. However, some letters that are structurally close have been differentiated in terms of sequence specificity: A is structurally close to N and T but does not belong to the same sequence cluster; C appears far from the other β-specific letter $[JMIBQ]$ in the sequential clustering. The amino-acid propensities of SA20-OMP thus provides additional information.

Use of SA20-OMP to rank models

As the structural alphabet is able to capture sequence/structure correlations, we experiment the ability of a scoring function based on it to assess the quality of structural models of OMPs. The usual tools for model evaluation are generally well suited for globular proteins but not for membrane proteins. We retrieved 47 structural models submitted by different predictor groups at the CASP3 experiment for the Omp32 protein. Some of these models have a correct barrel architecture, while others comprise many α helices. The sequence to be modeled is 332 residue long. Some groups predicted the structure only for parts of the sequence. Indeed, the model size varies from 13 to 332 residues. The Omp32 protein structure has since been deposited in the PDB with id 1e54. The quality of the models is measured by the Cα rmsd with the structure of 1e54. The sequence-to-structure adequacy is separately assessed on the 47 structural model thanks to a simple scoring function based on the sequence specificity of structural letters. Since 1e54 is part of our *OMPset*, it is removed from the data set before computing the probabilities needed to compute the parameters of our scoring function. We used pseudo-counts in order to avoid zero probabilities. It means that the counts are not initialized to zero but to a fix value (here, the same value is used for all probabilities). Best results were obtained with a pseudo-count equal to 1.3.

To see if our scoring function is able to correctly rank the model, we plot the scores against the rmsd. The resulting plot is shown on Figure 10a. It can be seen that the true structure obtains the maximum score. Moreover, there is a good correlation between adequacy scores and rmsd: models similar to the true structure have high scores while models far from the true structure have lower scores. Considering only the 32 models longer than 200 residues, we obtain a Pearson correlation coefficient equal to -0.74, with an associated p-value equal to 1.25e-06. These scores result from the sequence/structure correlation of 20 structural letters. Similar scores can be computed using the secondary structure instead of structural letters to describe the local structure of fragments. We computed adequacy scores using the STRIDE assignments, reduced into three classes (helix/strand/coil). The comparison between secondary structure scores and rmsd is shown on Figure 10b. In that case, the target structure has a score lower than some models with rmsd greater than 5 Å. The Pearson correlation coefficient in this case equals to -0.66 (p-value=3.37e-05).

This very promising result indicates that a scoring function based on SA20-OMP is very efficient to

correctly identify the correct structural models in a blind experiment context.

Discussion

It is now well established that OMP exhibit sequence specificity [12]. Thanks to our local alphabet, we question whether OMPs have original local structural features, which, to our knowledge, has never been done before.

Choosing the appropriate data sets

Few OMP structures are available in the PDB. In this study, we considered the maximal data set available without redundancy, which is a small data set of 17 structures. To assess the specificity of OMP structures compared to globular β proteins, a set of globular proteins was needed. In their study, Jackups and Liang used a set of 26 globular proteins with β-barrel like architecture, built by searching for structural homologs with OMP structures [28]. 13 of this globular proteins are β-barrel according to SCOP classification [53]. This data set contains only 42% of β-strand (*vs* 60% for the *OMPset*), thus it is not suitable for our comparison. We then considered a data set of PDB structures that are defined as β architecture by the CATH classification [54], excluding membrane proteins and filtered for sequence redundancy by the PISCES website [42]. The final set of 63 structures was only 39% β and then could not be used for comparison. For these reasons, we compiled a list of globular structures, the *GBset*, with high β-strand content (58%), regardless of their architecture. Nonetheless, we trained alphabets on the 26 globular structures used by Jackups and Liang, the 13 barrels of this data set, and the 63 β-barrels of CATH classification. These globular alphabets were significantly different from SA20-OMP. In particular, we observed the same tendency of structure compatibility toward the different alphabets.

Optimality of SA20-OMP

The size of the alphabet depends on the criterion used to chose the optimal size. Here, we obtain 20 structural letters using BIC. The Aikaike Information Criterion (AIC), another penalized log-likelihood criterion [55] selects a bigger alphabet. Indeed, AIC does not reach its maximum in the size range 1-25

structural letters (see supplementary data). This is not desirable, since a very large size of alphabet would lead to a poor parameter estimation. We could also use a structural criterion, such as the mean rmsd within each class ($rmsd_w$). For example, a threshold of 0.5Å would select a structural alphabet with 12 structural letters. In a 12 letters alphabet, four structural letters only specifically describe β-strands (see Figure 2). In this case, the alphabet would be a less accurate tool to study OMP structures. The BIC thus selects a model of both (i) a reasonable size, and consequently, a correct parameter estimation, and (ii) a number of structural letter that provides a detailed description of the structures.

Originality of SA20-OMP

We used the approach developed by Camproux et al [31, 34] that defined a generic alphabet, SA27. It was learned on 1,429 globular structures, resulting in a total data set 56,167 fragments [34]. This alphabet, composed of 27 structural letters is composed of 4 structural letters that specifically describe α-helices, 5 that specifically describe β-strands, the remaining 18 letters describing the loops. Here, we apply the same learning method on 17 OMP structures, representing a data set of 6,143 fragments, and obtain a 20 letter alphabet with 6 β-specific letters, 9 β-compatible and 5 are β-incompatible. The resulting alphabet is particularly stable, as assessed by a jackknife method. SA20-OMP, although smaller than the generic alphabet, has more β-structural letters. In terms of local fit approximation, SA20-OMP provides a mean $rmsd_w$ equal to 0.35 Å, which is to be compared to the local fit obtained with SA27 on globular structures, i. e., a mean $rmsd_w$ of 0.23 Å [34].

During this study, another structural alphabet, SA20-GB, was learned on a set of globular all-β structures, the *GBset*. *GBset* has been selected to have a similar β-strand content as *OMPset*. SA20-GB, like SA20-OMP, contains six β-specific letters. Its mean $rmsd_w$ is equal to 0.38 Å, thus similar to the mean $rmsd_w$ of SA20-OMP. Roughly, if the 7 most helical letters were excluded from SA27, the virtual resulting alphabet would have a mean $rmsd_w$ equal to 0.30 Å, a little lower than the mean $rmsd_w$ of SA20-OMP and SA20-GB. This small difference may be explained by the different amount of structures available to train the alphabets. Indeed, using a larger dataset could help to learn more precise structural letters.

When focusing on β-specific letters, the $rmsd_w$ varies from 0.21 Å to 0.31 Å for SA20-OMP, 0.25 Å and 0.33 Å in SA20-GB and 0.21 Å and 0.25 Å for SA-27, indicating in all cases the identification of very

precise β-specific letters.

SA20-OMP unravels the β-barrel architecture

A structural alphabet provides a description of the internal architecture of protein structures, in terms of overlapping fragments. Here, we show that the optimal description of β-barrel proteins can be achieved using 20 structural letters of 4 residues each. The transitions between structural letters are very constrained, as assessed by a global N^o_{eq} equal to 7 (far from 20 in case of all equal transitions. This expected number is to be compared to the N^o_{eq} obtained for the generalist structural alphabet AS27, equal to 8. In proportion, 8 transitions for 27 letter is lower than 7 transitions for 20 letters. This may be explained by the low helical content of OMP structures: the four most helical letters in SA27 display very high self-transitions and are indeed associated to very low N_{eq} values.

Albeit their relatively similar geometry, the six β-specific letters of SA20-OMP are well-distinct from each others, as shown by the $rmsd_b$. For example, some letters are more extended at the beginning (N and C), while others are more extended at the end (B and I). Moreover, some letters appear at the beginning of β-strands (J, B and Q), while others correspond the end of β-strands (I), as shown by the transition probabilities. Globally, the transitions bewtween β-specific letters [*JMBICQ*] are very constrained: mean N^o_{eq} is around 4. These constraints are stronger in SA20-OMP than in globular alphabets: mean N^o_{eq} equal respectively to 8 and 9 for the six β-specific letters of SA20-GB and the five β-specific letters of SA27. The possible pathways to form a β-strand is thus more limited in OMP than in globular structures. Roughly, to form a β-strand made of three structural letters, there are 96 possible pathways in SA20-OMP ($6 \times 4 \times 4$) versus 256 pathways in SA20-GB and 405 pathways in SA27.

SA20-OMP thus highlights the intrinsic organization of OMP structures thanks to a limited number of canonical fragments and captures subtle details of β-strand architecture. β-strands appear to be very constrained.

Local structure specificity of OMP structures

Some new features emerge from the analysis of OMP structures in SA20-OMP. The composition of β-strands, in terms of structural letters, is the same for all barrel size, but we found that different structural letters are preferred in different regions of the membrane. Letter J is more frequent in the periplasmic cap

region, while letters B and I are more frequent in the extracellular cap-region. It could be explained by the asymmetrical nature of the membrane. Previous study revealed amino-acid propensities in different regions of OMP structures [28]. In particular, the "positive outside rule" states that basic residues lysine and arginine are preferentially found in the extracellular cap region. Here we observe a preference of letter J in the extracellular cap region. These two findings seems to bring complementary information, since letter J does not specifically favor lysine or arginine residues. However, the structural tendencies identified in this study are in agreement with some of the propensities reported in [28]: letter I, over-represented in the periplasmic cap, is characterized by a strong preference for glycine at position 3, and glycine is over-represented in this region according to [28]. The comparison of different structural alphabets indicates that, despite the same secondary structure content, OMP structure have specific characteristics. The structural letters of SA20-OMP are more extended than the structural letters of SA20-GB. It seems that β-strands are more extended in membrane than in aqueous environment.

Specificity of fragment pairing

In their study, Jackups and Liang identified the amino-acid pairwise specificity in OMPs [28]. Using a non-parametric approach and a distinction between three kinds of inter-strand interactions they were able to discover pairwise interstrand motifs and anti-motifs.

With SA20-OMP, we found that there is also a structural pairwise specificity in OMP structures, more patent than what occurs in globular structures. Some of the amino-acid motifs identified in [28] are in agreement with the structural pairs identified in the present study. For instance, the motif leucine/tyrosine is in agreement with the preferred structural pair formed by structural letters M and B: letter M favors leucine at position 3 and letter B favors leucine at position 2 (when two structural letters are paired, their respective second and third Cα are facing). However, the inter-strand motifs and motifs presented by Jackups and Liang are identified using a background amino-acid frequency in β-strands whereas in Figure 8, the background frequency is the overall frequency in OMP sequences. A detailed comparison of Jackups and Liang's motifs with the amino-acid preference of structural letters compared to amino-acid frequency in β-strands is presented in supplementary data. For instance, the motif glycine/glutamine is coherent with the preferred pair formed by letters C and B that respectively favor glycine and glutamine residues at their second and third residues (remind that when two structural letters are paired, their

respective second and third Cα are facing). In the same way, motifs glycine/isoleucine, leucine/tyrosine, alanine/alanine and glycine/valine are respectively coherent with the preferred pairs *MI*, *MB*, *QQ* and *IC*.

Integration of sequential information

SA20-OMP is a structure-based classification, but sequential information can also been extracted. Fragment clustering shows that the sequential specificity is an additional information. When combined in a simple adequacy score, this specificity allows a very efficient discrimination of models that were submitted during CASP3 for the target structure 1e54A. This result open a very promising perspective of using our structural alphabet in a predictive framework. The transition between structural letters are governed by precise transition rules. Taking into account these transitions could help to obtain a better discrimination among structural models. Since the pairing of structural letters is not random, this information could also help. It requires, however, that the target as been identified as OMP, which is already an important scientific question [56].

Conclusion

Our alphabet, SA20-OMP, constitutes a new tool to specifically study the structure of β-barrel membrane proteins. For the first time, we show that OMP have specific local structural features and highly specific transition rules. The use of SA20-OMP also reveals a structural pairwise specificity between β-strands. We showed, on a case study, that SA20-OMP could be used to score structural models. Since the 3D prediction of membrane proteins is still a difficult task [57], we believe that such an approach could be of help for choosing the correct model among a collection. Another major application of SA20-OMP is that it allows a compression of 3D structures into 1D sequence of structural letters. Such representations are well suited for recurrent structural pattern analysis [37] and structure comparison, mining and prediction.

Acknowledgment

We would like to acknowledge Catherine Etchebest from EBGM for helpful discussions during the redaction, Leslie Regad from EBGM and François Rodolphe from MIG for critical reading of the manuscript.

We are grateful to INRA for awarding a Fellowship to JM.

References

[1] Marsden, R. L., Lee, D., Maibaum, M., Yeats, C. & Orengo, C. A. (2006). Comprehensive genome analysis of 203 genomes provides structural genomics with new insights into protein family space. *Nucleic Acids Res.* **34**, 1066-1080.

[2] Wimley, W. C. (2002). Toward genomic identification of beta-barrel membrane proteins: composition and architecture of known structures. *Protein Sci.* **1**1, 301-312.

[3] Schulz, G. E. (2000). beta-Barrel membrane proteins. *Curr. Opin. Struct. Biol.* **10**, 443-447.

[4] Berman, H. M., Westbrook, J., Feng, Z., Gilliland, G., Bhat, T. N., Weissig, H., Shindyalov, I. N. & Bourne, P. E. (2000). The Protein Data Bank. *Nucleic Acids Res.* **28**, 235-242.

[5] `http://blanco.biomol.uci.edu/Membrane_Proteins_xtal.html`.

[6] Schulz, G. E. (2002). The structure of bacterial outer membrane proteins. *Biochim. Biophys. Acta.* **1565**, 308-317.

[7] Broutin, I., Benabdelhak, H., Moreel, X., Lascombe, M. B., Lerouge, D., Ducruix, A. (2005). Expression, purification, crystallization and preliminary X-ray studies of the outer membrane efflux proteins OprM and OprN from Pseudomonas aeruginosa. *Acta. Crystallograph. Sect. F Struct. Biol. Cryst. Commun.* **1**, 315-318.

[8] Liu, Q., Zhu, Y., and Wang, B. & Li, Y. (2003). Identification of beta-barrel membrane proteins based on amino acid composition properties and predicted secondary structure. *Comput. Biol. Chem.* **27**, 355-361.

[9] Gromiha, M. M. & Suwa, M. (2003). Variation of amino acid properties in all-beta globular and outer membrane protein structures. *Int. J. Biol. Macromol.* **32**, 93-98.

[10] Gromiha, M. M. (2005). Motifs in outer membrane protein sequences: applications for discrimination. *Biophys. Chem.* **117**, 65-71.

[11] Gromiha, M. M., Ahmad, S. & Suwa, M. (2005). Application of residue distribution along the sequence for discriminating outer membrane proteins. *Comput. Biol. Chem.* **29**, 135-142.

[12] Gromiha, M. M. & Suwa, M. (2005). A simple statistical method for discriminating outer membrane proteins with better accuracy. *Bioinformatics.* **21**, 961-968.

[13] Martelli, P. L., Fariselli, P., Krogh, A. & Casadio, R. (2002). A sequence-profile-based HMM for predicting and discriminating beta barrel membrane proteins. *Bioinformatics.* **18**, W46-53.

[14] Liu, Q., Zhu, Y.-S., Wang, B.-H. & Li, Y.-X. (2003). A HMM-based method to predict the transmembrane regions of beta-barrel membrane proteins. *Comput. Biol. Chem.* **27**, 69-76.

[15] Bigelow, H. R., Petrey, D. S., Liu, J., Przybylski, D. & Rost, B. (2004). Predicting transmembrane beta-barrels in proteomes. *Nucleic Acids Res.* **32**, 2566-2577.

[16] Bagos, P. G., Liakopoulos, T. D., Spyropoulos, I. C. & Hamodrakas, S. J. (2004). A Hidden Markov Model method, capable of predicting and discriminating beta-barrel outer membrane proteins. *BMC Bioinformatics.* **5**, 29.

[17] Jacoboni, I., Martelli, P. L., Fariselli, P., De Pinto, V. & Casadio, R. (2001). Prediction of the transmembrane regions of beta-barrel membrane proteins with a neural network-based predictor. *Protein Sci.* **10**, 779-787.

[18] Gromiha, M. M., Ahmad, S. & Suwa, M. (2004). Neural network-based prediction of transmembrane beta-strand segments in outer membrane proteins. *J. Comput. Chem.* **25**, 762-767.

[19] Gromiha, M.M., Ahmad, S. & Suwa, M. (2005). TMBETA-NET: discrimination and prediction of membrane spanning beta-strands in outer membrane proteins. *Nucleic Acids Res.* **33**, S164-167.

[20] Bigelow, H. & Rost, B. (2006). PROFtmb: a web server for predicting bacterial transmembrane beta barrel proteins. *Nucleic Acids Res.* **34**, S186-188.

[21] Natt, N. K., Kaur, H. & Raghava, G. P. S. (2004). Prediction of transmembrane regions of beta-barrel proteins using ANN- and SVM-based methods. *Proteins.* **56**, 11-18.

[22] Park, K.-J., Gromiha, M. M., Horton, P. & Suwa, M. (2005). Discrimination of outer membrane proteins using support vector machines. *Bioinformatics.* **21**, 4223-4229.

[23] Garrow, A. G., Agnew, A. & Westhead, D. (2005). TMB-Hunt: an amino acid composition based method to screen proteomes for beta-barrel transmembrane proteins. *BMC Bioinformatics*. **6**, 56.

[24] Bagos, P. G., Liakopoulos, T. D. & Hamodrakas, S. J. (2005). Evaluation of methods for predicting the topology of beta-barrel outer membrane proteins and a consensus prediction method. *BMC Bioinformatics*. **6**, 7.

[25] Gromiha, M. M. & Suwa, M. (2006). Influence of amino acid properties for discriminating outer membrane proteins at better accuracy. *Biochim. Biophys. Acta*. **1764**, 1493-1497.

[26] Waldispuhl, J., Berger, B., Clote, P. & Steyaert, J.-M. (2006). Predicting transmembrane beta-barrels and interstrand residue interactions from sequence. *Proteins*. **65**, 61-74.

[27] Seshadri, K., Garemyr, R., Wallin, E., von Heijne, G. & Elofsson, A. (1998). Architecture of beta-barrel membrane proteins: analysis of trimeric porins. *Protein Sci*. **7**, 2026-2032.

[28] Jackups, R. & Liang, J. (2005). Interstrand pairing patterns in beta-barrel membrane proteins: the positive-outside rule, aromatic rescue, and strand registration prediction. *J. Mol. Biol*. **354**, 979-993.

[29] Jackups, R., Cheng, S. & Liang, J. (2006). Sequence Motifs and Antimotifs in beta-Barrel Membrane Proteins from a Genome-Wide Analysis: The Ala-Tyr Dichotomy and Chaperone Binding Motifs. *J. Mol. Biol*. **363**, 611-623.

[30] Park, B. H., & Levitt, M. (1995). The complexity and accuracy of discrete state models of protein structure. *J. Mol. Biol.* **249**, 493-507.

[31] Camproux, A. C., Tufféry, P., Chevrolat, J. P., Boisvieux, J. F. & Hazout, S. (1999). Hidden Markov model approach for identifying the modular framewor k of the protein backbone. *Protein Eng*. **12**, 1063-1073.

[32] de Brevern, A. G., Etchebest, C. & Hazout, S. (2000). Bayesian probabilistic approach for predicting backbone structures in terms of protein blocks. *Proteins*. **41**, 271-287.

[33] Kolodny, R., Koehl, P., Guibas, L. & Levitt, M. Small libraries of protein fragments model native protein structures accurately. *J. Mol. Biol*. **323**, 297-307.

[34] Camproux, A.C., Gautier, R. & Tufféry, P. (2004). A hidden markov model derived structural alphabet for proteins. *J. Mol. Biol.* **339**, 591-605.

[35] Guyon, F., Camproux, A.-C., Hochez, J. & Tufféry, P. (2004). SA-Search: a web tool for protein structure mining based on a Structural Alphabet. *Nucleic Acids Res.* **32**, W545-548.

[36] Tyagi, M., Gowri, V. S., Srinivasan, N., de Brevern, A. G. & Offmann, B. (2006). A substitution matrix for structural alphabet based on structural alignment of homologous proteins and its applications. *Proteins.* **65**, 32-39.

[37] Regad, L., Martin, J. & Camproux, A.C. (2006). Identification of non Random Motifs in Loops Using a Structural Alphabet. *Proceedings of IEEE Symposium on Computational Intelligence in Bioinformatics and Computational Biology.* 92-100.

[38] Etchebest, C., Benros, C., Hazout, S. & de Brevern, A. G. (2005). A structural alphabet for local protein structures: improved prediction methods. *Proteins.* **59**, 810-827.

[39] Camproux, A. C. & Tufféry, P. (2005). Hidden Markov model-derived structural alphabet for proteins: the learning of protein local shapes captures sequence specificity. *Biochim. Biophys. Acta.* **1724**, 394-403.

[40] Frishman, D. & Argos, P. (1995). Knowledge-based protein secondary structure assignment. *Proteins.* **23**, 566-579.

[41] Martin, J., Letellier, G., Marin, A., Taly, J.-F., de Brevern, A. G. & Gibrat, J.-F. (2005). Protein secondary structure assignment revisited: a detailed analysis of different assignment methods. *BMC Struct. Biol.* **5**, 17.

[42] Wang, G. & Dunbrack, R. L. (2003). PISCES: a protein sequence culling server. *Bioinformatics.* **19**, 1589-1591.

[43] Alland, C., Moreews, F., Boens, D., Carpentier, M., Chiusa, S., Lonquety, M., Renault, N., Wong, Y., Cantalloube, H., Chomilier, J., Hochez, J., Pothier, J., Villoutreix, B. O., Zagury, J.-F. & Tufféry, P. (2005). RPBS: a web resource for structural bioinformatics. *Nucleic Acids Res.* **33**, W44-49.

[44] Moult J., Hubbard T., Fidelis K. & Pedersen J.T. (1999). Critical assessment of methods of protein structure prediction (CASP): round III. *Proteins*. **Suppl 3**, 2-6.

[45] Castrignano T., De Meo P.D., Cozzetto D., Talamo I.G. & Tramontano A. (2006). The PMDB Protein Model Database. *Nucleic Acids Res*. **34**, D306-309.

[46] Zeth K., Diederichs K., Welte W., & Engelhardt H. (2000) Crystal structure of Omp32, the anion-selective porin from Comamonas acidovorans, in complex with a periplasmic peptide at 2.1 A resolution. *Structure*. **9**, 981-992.

[47] Rabiner, L. R. (1989). A tutorial on Hidden Markov Models and selected applications in speech recognition. *Proceedings of the IEEE*. **77**, 257-286.

[48] Schwartz, G. (1978). Estimating the dimension of a model. *Annals of Statistics*. **6**, 461-464.

[49] Regad, L., Guyon, F., Maupetit, J., Tufféry, P. & Camproux A.C. (2006) A Hidden Markov Model Applied to the Protein 3D Structure Analysis. *Computational Statistics Data and Analysis*. In press.

[50] Lomize, M. A., Lomize, A. L., Pogozheva, I. D. & Mosberg, H. I. (2006). OPM: orientations of proteins in membranes database. *Bioinformatics*. **22**, 623-625.

[51] Lomize, A. L., Pogozheva, I. D., Lomize, M. A. & Mosberg, H. I. (2006). Positioning of proteins in membranes: a computational approach. *Protein Sci*. **15**, 1318-1333.

[52] Sander, O., Sommer, I., Lengauer, T. (2006) Local protein structure prediction using discriminative models. *BMC Bioinformatics*. **7**, 14.

[53] Lo Conte, L., Ailey, B., Hubbard, T. J., Brenner, S. E., Murzin, A. G. & Chothia, C. (2000). SCOP: a structural classification of proteins database. *Nucleic Acids Res*. **28**, 257-259.

[54] Orengo, C. A., Michie, A. D.,Jones, S., Jones, D. T., Swindells, M. B. & Thornton, J. M. (1997). CATH–a hierarchic classification of protein domain structures. *Structure*. **5**, 1093-1108.

[55] Akaike H. (1973). Information theory and Extension of the Likelihood Ratio Principe. *Proceedings of the second International Symposium of Information theory*. 257-281.

[56] Gromiha, M. M., Yabuki, Y., Kundu, S., Suharnan, S., Suwa. M. (2007) TMBETA-GENOME: database for annotated beta-barrel membrane proteins in genomic sequences. *Nucleic Acids Res.* **35**, D314-316.

[57] Elofsson A. & von Heijne G. (2007) Membrane Protein Structure: Prediction vs Reality. *Annu Rev Biochem.* In press.

Table 1: Description of the *OMPset*. *PDBcode* corresponds to the PDB code (4 letters) and the fifth character, if any, corresponds to the protein chain. *Nstrands* is the number of strands involved in the barrel. *Lstrands* denotes the mean length of β-strands, with associated standard deviation between parentheses. Strands are assigned by STRIDE, except 1kmo (noted by *) which is assigned by KAKSI as STRIDE does provides any assignment for this protein. The *biological unit* is taken from the PDB file.

PDBcode	Nstrands	Lstrands	Biological unit
1bxw	8	17.1(3.8)	monomer
1qj8	8	18.9(3.1)	trimer
1p4t	8	17.1(2.1)	monomer
1k24	10	19.1(3.8)	monomer
1i78A	10	26.2(2.3)	monomer
1qd6C	12	15.4(3.0)	monomer
1uyn	12	19.3(3.0)	monomer
2por	16	13.7(3.3)	trimer
1prn	16	12.9(2.6)	trimer
2omf	16	15.6(3.3)	trimer
1e54	16	14.4(3.4)	monomer
2mprA	18	16.7(3.5)	trimer
1a0sP	18	15.3(1.9)	trimer
1fep	22	16.3(3.9)	monomer
2fcp	22	18.1(5.1)	monomer
1kmo	22	16.5*(2.6)	monomer
1nqe	22	16.5(2.6)	monomer
global	256	16.6(4.2)	

Table 2: Description of the 20 structural letters of SA20-OMP. Structural letters are ranked according to the value of d_2. Nb and % are the frequencies and relative frequencies of each structural letter in the *OMPset*. d_1, d_2, d_3 and d_4 denote the mean descriptors associated to each letter. The symbol * indicates the descriptor with the highest standard deviation. rmsd_w is the average rmsd within the fragments encoded by a given structural letter. rmsd_b is the *minimum* average rmsd observed between the considered letter and all the others, obtained for the letter appearing between parentheses. $\%\beta_s$ and $\%\beta_k$ denote respectively the fraction of a given structural letter that correspond to a strand conformation assigned by STRIDE and KAKSI in the *OMPset*.

State	Nb	%	d_1 (Å)	d_2 (Å)	d_3 (Å)	d_4 (Å)	rmsd_w (Å)	rmsd_b (Å)	$\%\beta_s$	$\%\beta_k$
A	323	5.26	5.45	5.35*	5.59	2.92	0.25	0.46 (*T*)	0.0	0.31
N	247	4.02	6.05	5.51	5.69	0.68*	0.44	0.75 (*A*)	2.3	3.7
T	132	2.15	5.47	6.23*	5.50	3.61	0.31	0.46 (*A*)	2.6	0.76
H	142	2.31	5.56	6.92*	5.59	-3.23	0.65	0.83 (*D*)	0.78	0.70
E	197	3.21	5.67	7.76	7.00	-0.40*	0.71	0.93 (*S*)	47.0	21.0
S	215	3.50	5.71	7.85*	6.84	2.65	0.40	0.80 (*O*)	52.0	33.0
D	162	2.64	6.77	8.25*	5.44	-3.50	0.38	0.61 (*F*)	2.8	8.0
O	278	4.53	5.64	8.45	6.25	2.75*	0.48	0.80 (*S*)	20.0	18.0
R	211	3.43	6.88	8.45*	6.35	-2.74	0.55	0.69 (*D*)	32.0	37.0
G	142	2.31	6.93	8.74	6.41	1.24*	1.03	0.94 (*R*)	42.0	43.0
F	259	4.22	6.76	9.22*	6.43	-3.24	0.34	0.61 (*D*)	32.0	41.0
P	276	4.49	6.08	9.23	5.93	0.11*	0.45	0.58 (*J*)	44.0	34.0
K	239	3.89	6.68	9.25	5.75	-1.84*	0.37	0.59 (*L*)	46.0	52.0
L	347	5.65	6.52	9.78	6.63	-1.97*	0.33	0.48 (*J*)	54.0	61.0
J	632	10.29	6.61	10.05	6.60	-0.60*	0.21	0.30 (*M*)	96.6	96.7
M	595	9.69	7.00	10.25	6.60	-0.79*	0.23	0.30 (*J*)	98.1	98.7
B	614	10.00	6.58	10.27	7.01	-0.10*	0.24	0.32 (*Q*)	96.9	97.7
I	324	5.27	6.64	10.30	7.20	0.96*	0.31	0.41 (*B*)	93.0	89.0
C	239	3.89	7.21	10.31	6.75	-2.01*	0.29	0.40 (*M*)	95.5	94.1
Q	569	9.26	7.00	10.57	6.98	-0.20*	0.24	0.32 (*B*)	97.4	98.2
total	6143						0.35		61.8	60.8

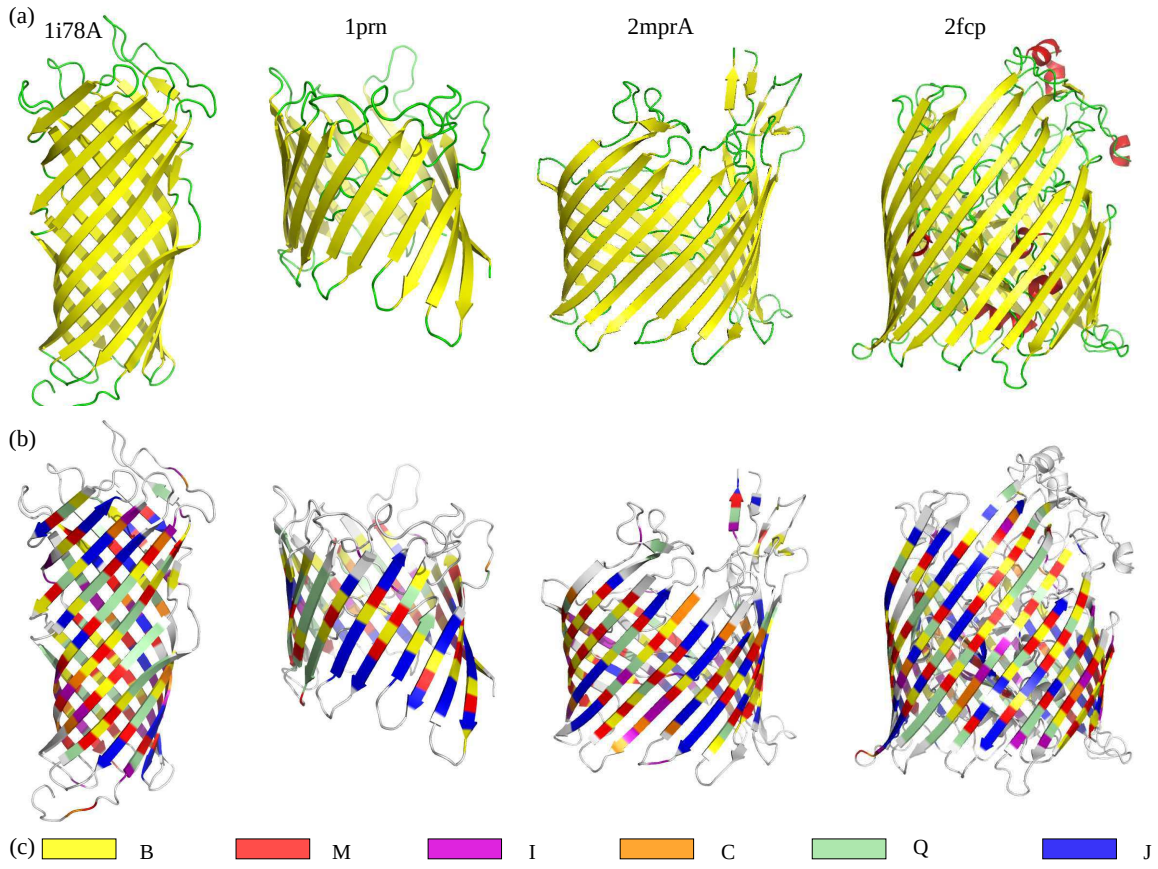


Figure 1: 3D structures of four OMPs with various barrel size and strand length. (a): structures colored according to the secondary structure assigned by KAKSI (yellow: β-strands, red: α-helix, green: coil), (b): structures colored according to the OMP-specific alphabet encoding, with correspondence between a structural letter and its third Cα, (c): color scheme used in (b). Only six structural letters out of 20, the β-specific letters [*BMICQJ*], are colored. From left to right: chain A of outer membrane protease OmpT from *Escherichia coli*, PDB code 1i78 (10 strands, mean strand length 26.2 residues); porin from *Rhodopseudomonas blastica*, PDB code 1prn (16 strands, mean strand length equal to 12.9 residues); chain A of maltoporin from *Salmonella typhimurium*, PDB code 2mpr (18 strands, mean strand length equal to 16.7 residues) and ferric hydroxamate uptake receptor Fuha from *Escherichia coli*, PDB code 2fcp (22 strands, mean length 18 residues, highly variable strand length). Images are generated using PyMOL (Warren L. DeLano, The PyMOL Molecular Graphics System,DeLano Scientific LLC, San Carlos, CA, USA. http://www.pymol.org).

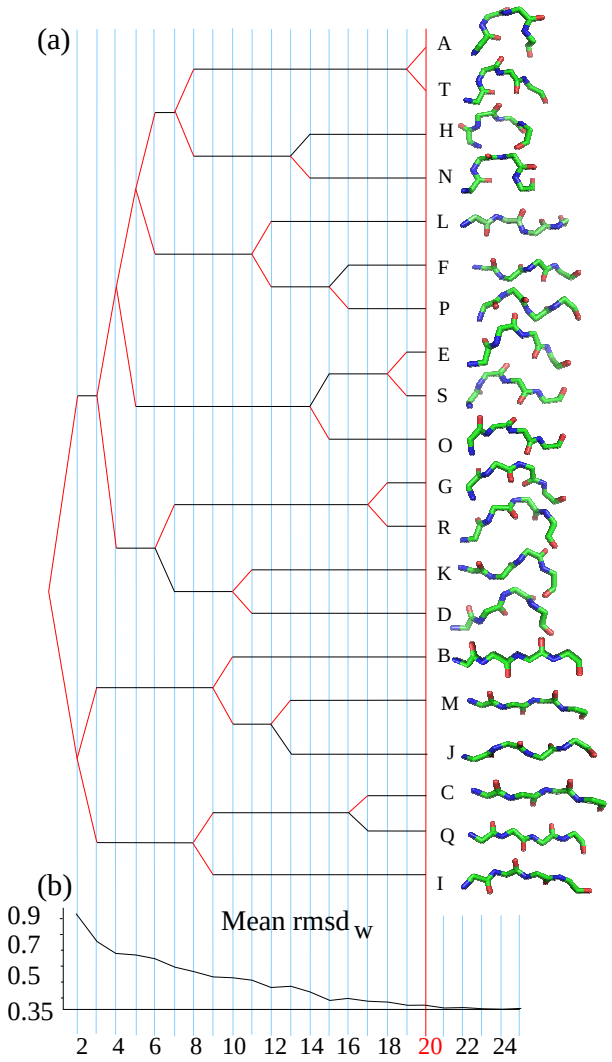
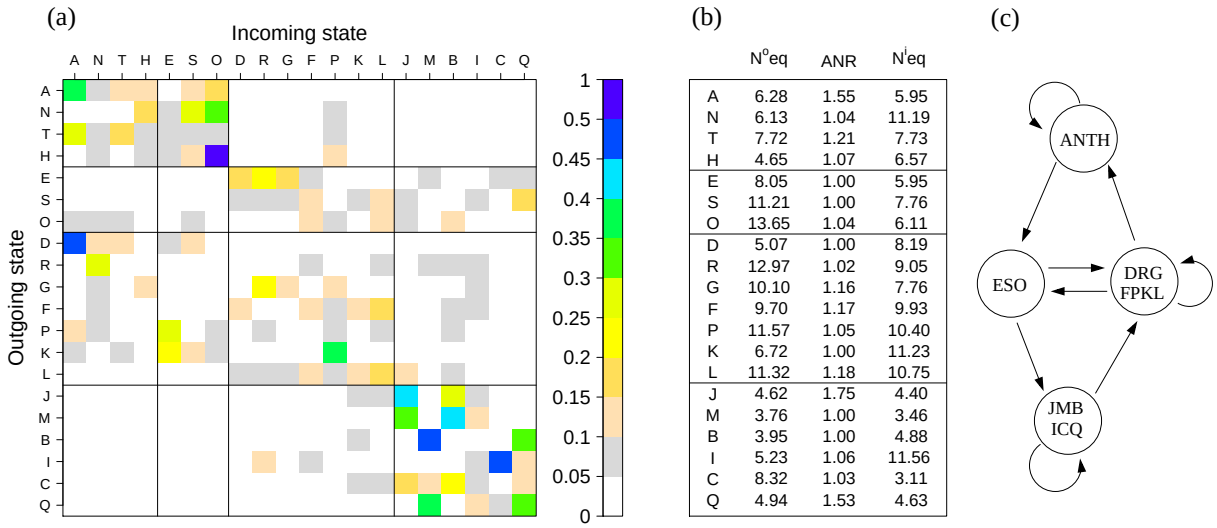


Figure 2: Schematic genesis of SA20-OMP. (a): splitting events and 3D representation of the 20 structural letters of the final alphabet. (b): global mean rmsd_w observed for the alphabet of corresponding size. Figures of structural letters are generated using PyMOL.



	N^o_{eq}	ANR	N^i_{eq}
A	6.28	1.55	5.95
N	6.13	1.04	11.19
T	7.72	1.21	7.73
H	4.65	1.07	6.57
E	8.05	1.00	5.95
S	11.21	1.00	7.76
O	13.65	1.04	6.11
D	5.07	1.00	8.19
R	12.97	1.02	9.05
G	10.10	1.16	7.76
F	9.70	1.17	9.93
P	11.57	1.05	10.40
K	6.72	1.00	11.23
L	11.32	1.18	10.75
J	4.62	1.75	4.40
M	3.76	1.00	3.46
B	3.95	1.00	4.88
I	5.23	1.06	11.56
C	8.32	1.03	3.11
Q	4.94	1.53	4.63

Figure 3: Transition rules of SA20-OMP. (a): graphical representation of transition probabilities. Probability ranges are indicated by different colors, from white for probabilities less than 0.05, to dark blue for probabilities close to 1. (b): number of equivalent outputs (N^o_{eq}), average number of repeats (ANR) and number of equivalent inputs (N^i_{eq}) of each structural letter. (c): main transitions between the four subgroups.

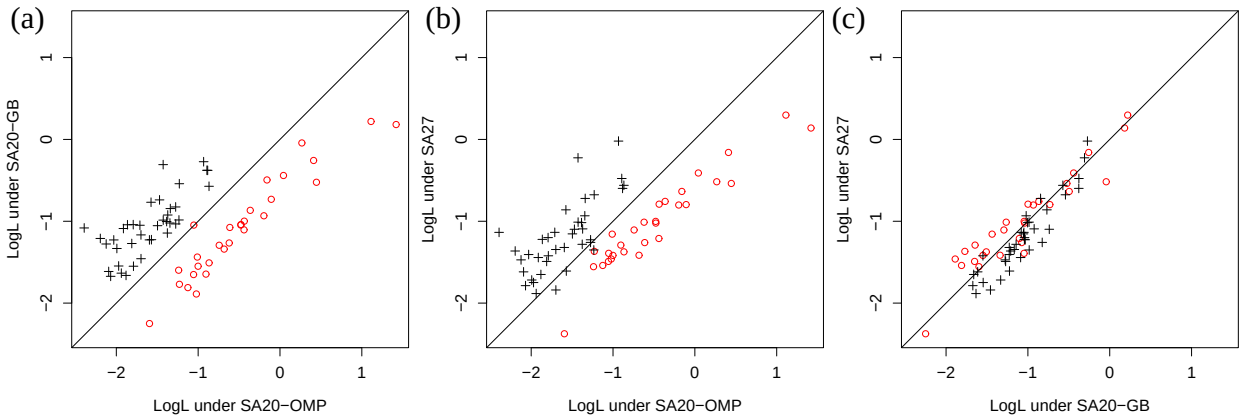


Figure 4: Comparison of log-likelihoods (LogL) of structures under different alphabets. Log-likelihoods are normalized by sequence lengths. Open red circles represent *OMPset* structures and black crosses represent *GBset* structures. (a): comparison between SA20-OMP and SA20-GB, (b): comparison between SA27 and SA20-OMP, (c): comparison between SA27 and SA20-GB.

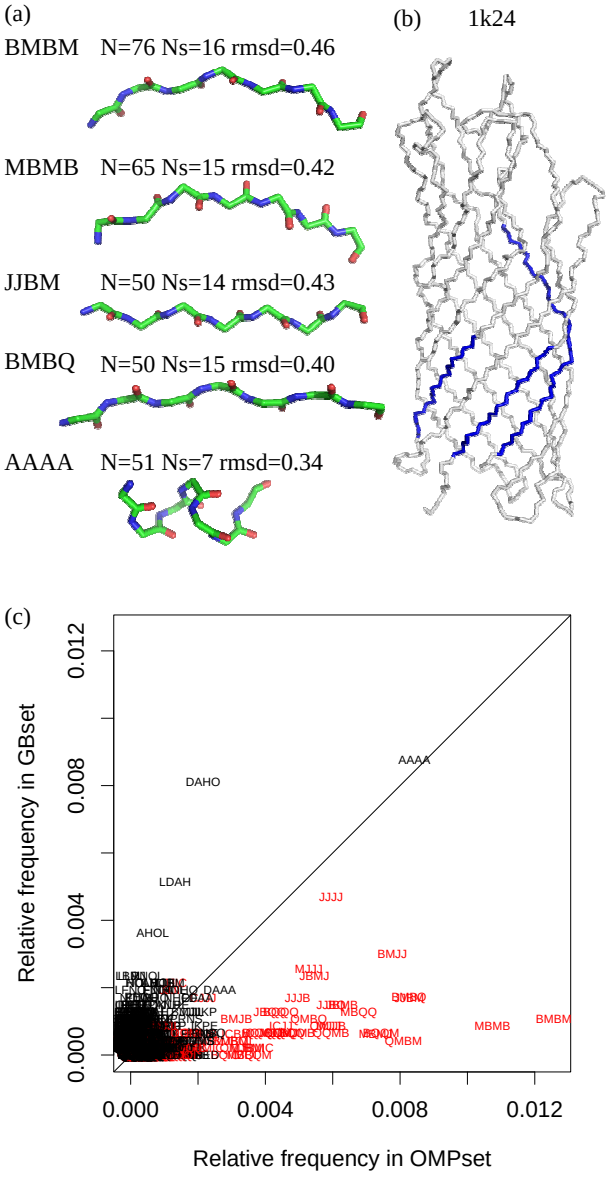


Figure 5: Recurrent structural patterns in OMP structures. (a): 3D representation of the five most frequent patterns. The associated rmsd are reported in Å. *N* and *Ns* denote respectively the frequency of the pattern and the number of protein structures that contain the pattern. (b): structural pattern *JJMB* in the structure 1k24. The structural pattern *JJMB* is highlighted in blue. (c) comparison of four letter pattern relative frequencies in *OMPset* and *GBset*. Red patterns are composed of β-specific letters. Images are generated by PyMOL.

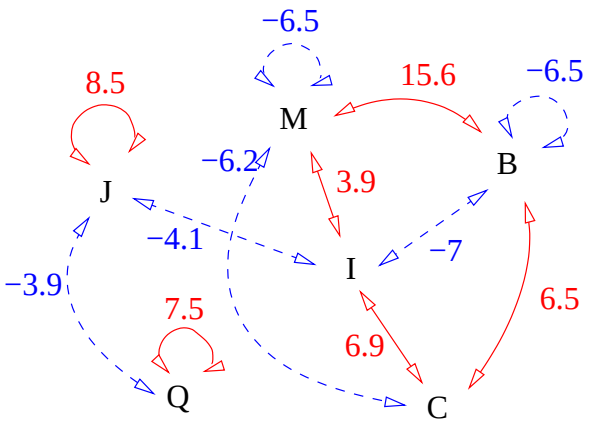


Figure 6: Preferred and avoided contacts between β-specific structural letters in the β-sheets in *OMPset*. Red and blue arrows indicate respectively significant over and under-represented pairs, with the associated Z-score.

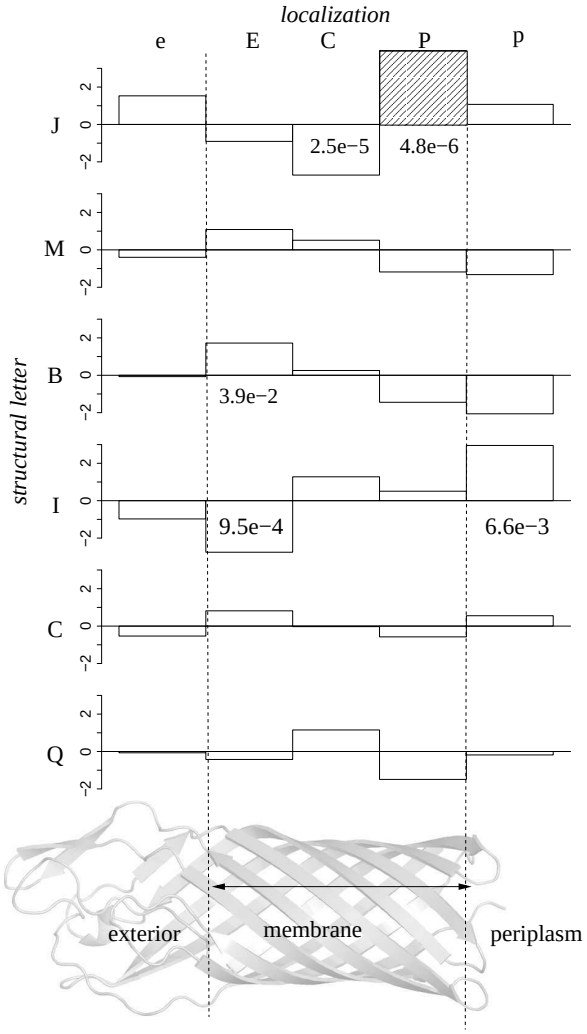


Figure 7: Propensity Z-scores for the structural letters [*JMBICQ*] in five distinct regions of the OMP structures. The Bonferroni-corrected level of significance in this case equals to 3.0. The significant Z-score is indicated by hatching. Significant p-values obtained with the non-parametric approach [28] are indicated on the plot.

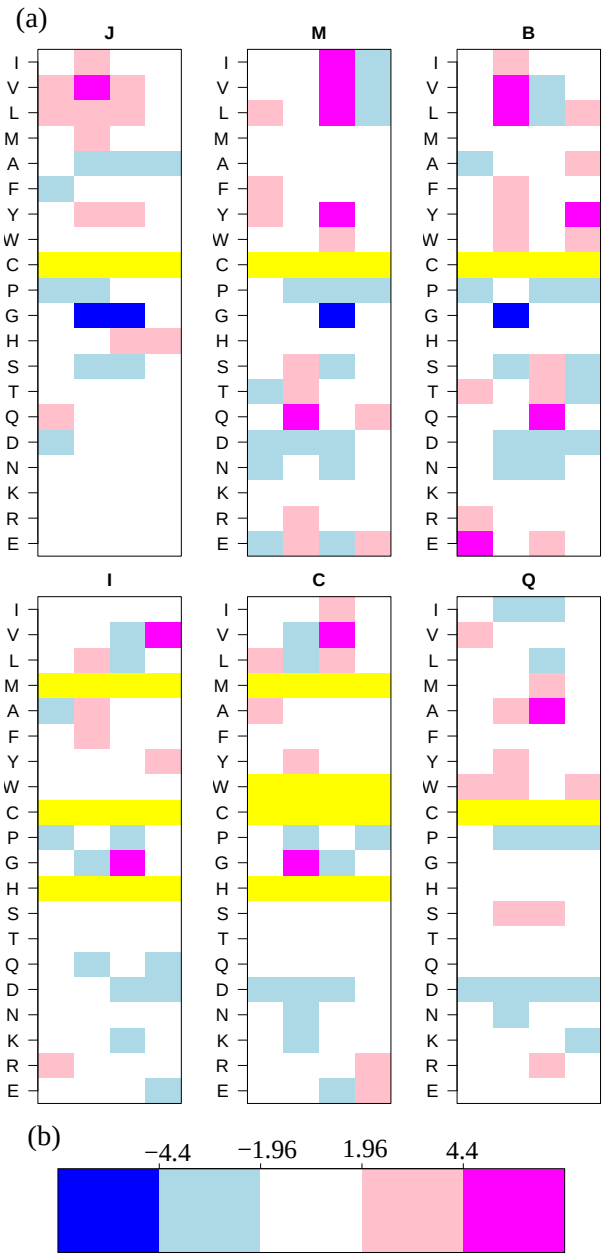


Figure 8: Z-scores of the 20 amino-acids in the four positions of the β-specific structural letters. (a): graphical representations of Z-scores, (b): color scale of Z-scores. Yellow indicates that the expected frequency is too low (less than five) to compute a Z-score.

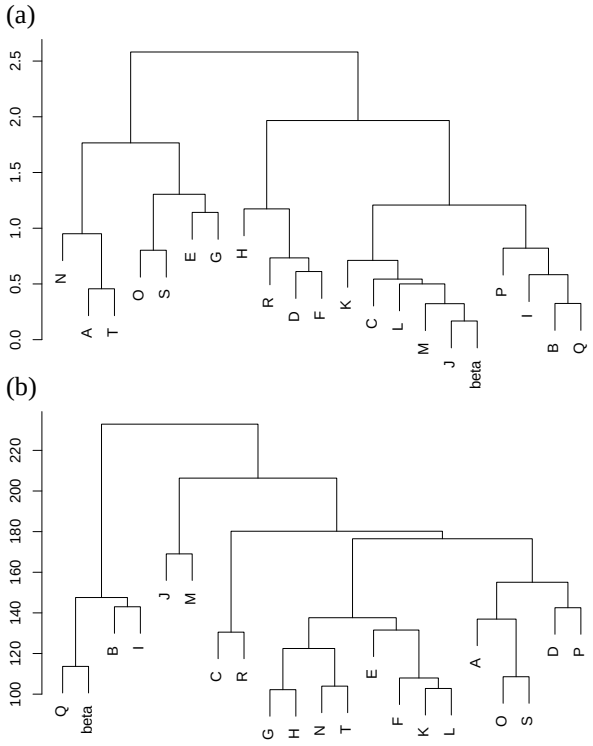


Figure 9: Hierarchical clustering of the 20 structural letters of SA20-OMP. (a): structural clustering using rmsd_b. (b): clustering according to amino-acid Z-scores distances. *beta* refers to the representative fragment of β-strands according to STRIDE. The complete linkage is used.

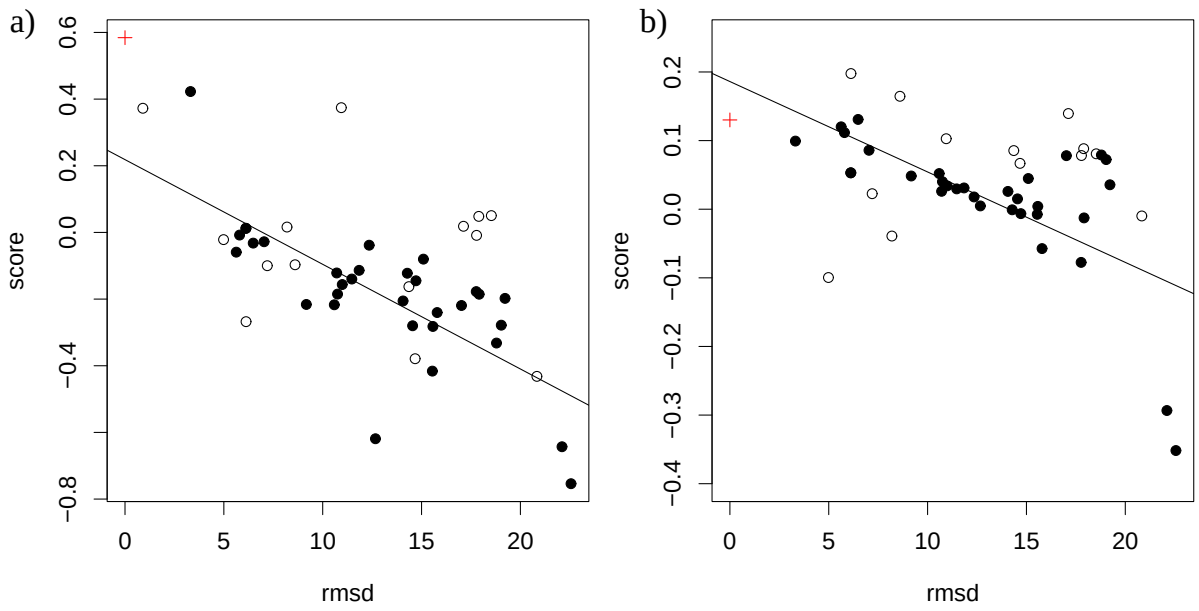


Figure 10: Correlation between adequacy scores and rmsd with the target, on models submitted for 1e54A at CASP3. a) scores based on SA20-OMP, b) scores based on STRIDE assignment. Open circles: models shorter than 200 residues, plain circles: models longer than 200 residues, red cross: structure of the target 1e54A.